



Michele Roggia, Benito Natale, Sandro Cosconati
Dipartimento di Scienze e Tecnologie Ambientali,
Biologiche e Farmaceutiche (DiSTABiF)
Università della Campania 'Luigi Vanvitelli', Caserta
sandro.cosconati@unicampania.it

<http://dx.medra.org/10.17374/CI.2026.108.3.48>

AI E SPAZIO CHIMICO: IL METODO PyRMD

L'intelligenza artificiale rivoluziona la scoperta di farmaci rendendo accessibile uno spazio chimico infinito. Il contributo descrive PyRMD, tool automatizzato di screening virtuale basato sull'algoritmo Random Matrix Discriminant. Si illustrano l'integrazione con il docking (PyRMD2Dock), la suite PyRMD Studio e la scoperta di un modulatore per il recettore GPR183.

Introduzione: l'intelligenza artificiale e la vastità dello spazio chimico

La ricerca e lo sviluppo di nuovi agenti terapeutici (*drug discovery*) affrontano da decenni una sfida sistemica: un progressivo aumento dei costi, associato a un elevato tasso di logoramento (*attrition rate*) nelle fasi di sviluppo clinico. La radice di questo problema risiede spesso nell'incapacità di identificare molecole con profili di efficacia e sicurezza ottimali sin dalle primissime fasi della ricerca. La soluzione a questa limitazione richiede l'esplorazione di porzioni inesplorate del cosiddetto "spazio chimico", definito come l'insieme di tutte le possibili molecole organiche compatibili con i criteri di "drug-likeness" (ad esempio, le regole di Lipinski). Le stime teoriche indicano che lo spazio chimico contenga un numero di molecole nell'ordine di 10^{60} . Di contro, le librerie storiche di composti sintetizzati fisicamente si fermano a pochi milioni di unità. Sebbene i recenti database virtuali *make-on-demand* (come le librerie REAL di Enamine) abbiano superato la soglia dei 10 miliardi di composti, esplorare queste librerie mediante metodologie computazionali tradizionali o test *in vitro* su scala *high-throughput* risulta tecnicamente ed economicamente insostenibile.

In questo scenario, l'Intelligenza Artificiale (AI) ha assunto un ruolo trasformativo permettendo di sviluppare modelli predittivi in grado di dedurre relazioni non lineari complesse tra la struttura chimica di una molecola e la sua attività biologica. Inoltre, i modelli di Machine Learning (ML) e Deep Learning (DL) consentono di bypassare l'eshaustività computazionale del calcolo basato sulle leggi della fisica

classica, fungendo da filtri ad altissima efficienza. Grazie a queste caratteristiche, l'AI non si limita a processare dati più velocemente, ma permette di navigare lo spazio chimico infinito in modo direzionale, identificando regioni arricchite di potenziali candidati farmaci (*hit compounds*).

Il fondamento matematico: l'algoritmo Random Matrix Discriminant

Nell'ambito dello screening virtuale basato sul ligando (Ligand-Based Virtual Screening, LBVS), l'approccio predominante si è a lungo basato sul calcolo della similarità molecolare (es. indice di Tanimoto su *fingerprint* vettoriali). Sebbene rapido, questo metodo fallisce nell'identificare composti attivi con scaffold chimici innovativi (*scaffold hopping*).

Per superare queste limitazioni, l'algoritmo *Random Matrix Discriminant* (RMD) introduce un approccio radicalmente diverso, fondato sulla teoria delle matrici casuali. L'assunto di base dell'RMD è che i set di dati chimici, specialmente quelli estratti da *repository* pubblici eterogenei, siano affetti da un elevato grado di rumore statistico. L'algoritmo si propone di isolare il "segnale" farmacoforico (l'insieme delle caratteristiche steriche, elettroniche e topologiche invarianti responsabili dell'interazione target-specifica) dal rumore di fondo generato da porzioni molecolari non essenziali per il legame. Operativamente, l'RMD analizza i descrittori molecolari calcolati per una serie di composti noti, sia attivi che inattivi. Attraverso un'analisi spettrale della matrice di covarianza associata a questi dati, l'algoritmo identifica autovalori e autovettori che definiscono lo spazio delle caratteristiche chimi-

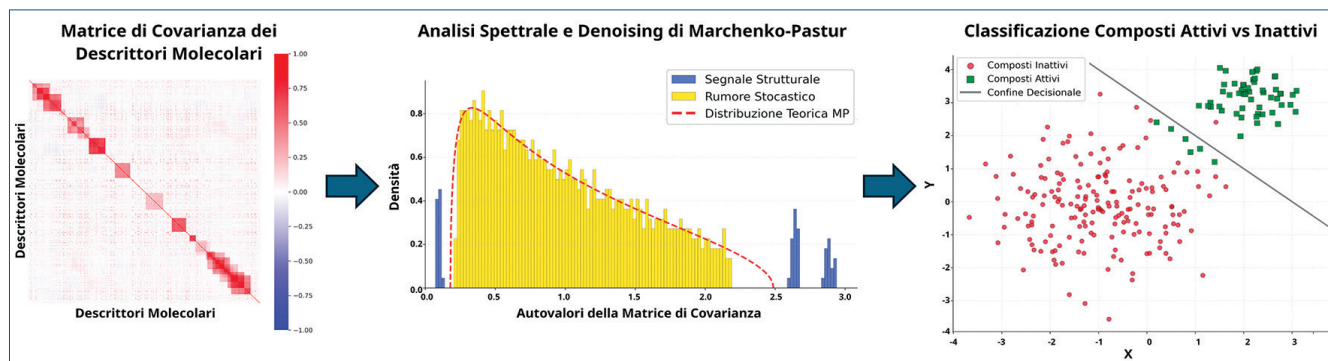


Fig. 1 - Rappresentazione schematica dell'algoritmo Random Matrix Discriminant (RMD)

che. Applicando soglie derivate dalla distribuzione di Marchenko-Pastur, l'RMD filtra le componenti attribuibili a fluttuazioni stocastiche casuali. Studi sistematici hanno dimostrato che questo approccio di *de-noising* conferisce all'RMD prestazioni predittive superiori, in termini di Area Under the ROC Curve (AUC) e Fattore di Arricchimento (EF), rispetto a complessi algoritmi di deep learning, come le reti neurali basate su grafi, garantendo al contempo una maggiore interpretabilità del modello chimico. Nella Fig. 1 il pannello illustra il processo di de-noising statistico applicato alla matrice di covarianza dei descrittori molecolari. Mediante l'analisi spettrale degli autovalori derivati dalla matrice di covarianza in accordo con la distribuzione di Marchenko-Pastur, l'algoritmo separa le fluttuazioni stocastiche (rumore di fondo) dalle componenti principali significative. Questo filtraggio dimensionale permette di estrarre esclusivamente il "segnale" farmacoforico invariante, essenziale per

la corretta discriminazione strutturale tra molecole attive e composti inattivi.

Automazione del flusso di lavoro: Il tool PyRMD

La complessità matematica dell'RMD ne ha inizialmente limitato l'uso a esperti di programmazione statistica. Per rendere la tecnologia fruibile alla più ampia comunità di chimici farmaceutici, è stato sviluppato PyRMD, uno strumento *open-source* interamente programmato in Python.

PyRMD risolve il collo di bottiglia della preparazione dei dati. Il software automatizza la procedura di recupero delle molecole dal database ChEMBL, estraendo ligandi attivi ed inattivi (anche questi ultimi impiegati nel processo di addestramento del modello) per un target indicato dall'utente e curando chimicamente le strutture (neutralizzazione delle cariche, rimozione dei sali, calcolo della stereochimica). Successivamente, converte i composti in vettori di *fingerprint* molecolari tramite

librerie come RDKit, istruendo automaticamente il classificatore RMD. Attraverso la modalità di *Benchmark*, il modello viene dunque sottoposto a rigorosi cicli di *cross-validation* (*repeated stratified K-fold*) per accertarne la robustezza predittiva, prima di procedere allo screening virtuale ad altissima velocità su librerie esterne composte da milioni di incognite (Fig. 2).

Screening ibrido e ottimizzazione delle risorse: PyRMD2Dock

Il limite intrinseco delle metodologie LBVS è la cecità rispetto alla conformazione tridimensionale del sito

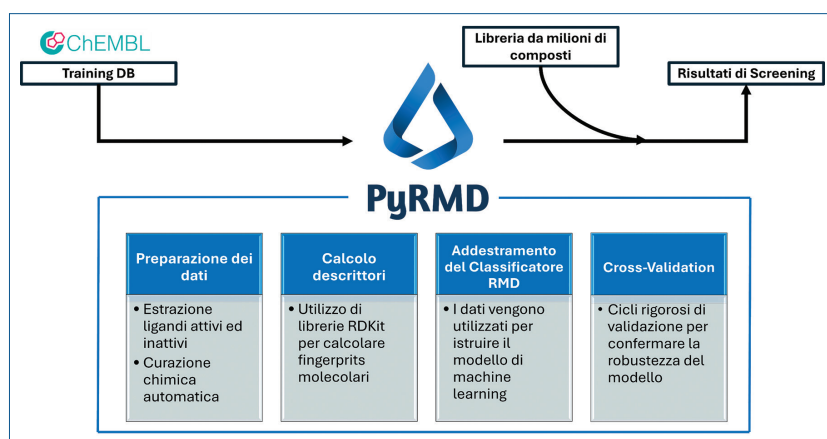


Fig. 2 - Workflow automatizzato di PyRMD (LBVS). Il diagramma illustra l'integrazione con il database ChEMBL, tutte le procedure automatizzate dal tool e lo screening di librerie virtuali di milioni di composti

di legame recettoriale. Per conferire robustezza strutturale alle predizioni, è stato strutturato il protocollo gerarchico PyRMD2Dock, che fonde la velocità dell'AI con il rigore fisico dello screening virtuale basato sulla struttura (Structure-Based Virtual Screening, SBVS).

Il flusso computazionale di PyRMD2Dock prevede un'architettura ibrida e integrata. Inizialmente, il protocollo utilizza AutoDock-GPU su una frazione casuale della libreria (es. il 5% del totale) per generare i dati necessari ad addestrare l'algoritmo di intelligenza artificiale PyRMD. Successivamente, PyRMD scansiona il resto della libreria chimica, operando una drastica riduzione dimensionale: il classificatore scarta un numero cospicuo di molecole, conservando esclusivamente quelle con la probabilità statistica più alta di presentare una buona affinità di legame.

Il sottoinsieme ristretto dei migliori candidati viene poi processato mediante docking molecolare finale. Sfruttando le unità di calcolo parallelo delle moderne schede grafiche (accelerazione CUDA), AutoDock-GPU simula le interazioni spaziali e termodinamiche tra il ligando flessibile e il target proteico.

I test di validazione condotti su 5 target farmacologici di rilevanza clinica, tra cui i recettori ADO-RA-2A e VEGFR2, hanno dimostrato che il protocollo PyRMD2Dock abbatte drasticamente i tempi di calcolo, rendendo fattibile lo screening di librerie "ultra-large" con risorse limitate. Le molecole valutate dal modello di intelligenza artificiale con il punteggio di confidenza più alto (RMD Score) presentano un netto arricchimento di composti ad alta affinità rispetto a una selezione puramente casuale (Fig. 3). Nella Fig. 3 lo schema illustra l'architettura impiegata per lo screening virtuale di librerie chimiche massive. Il processo ha inizio con una fase di addestramento in cui un campione casuale di molecole viene processato tramite AutoDock-GPU per istruire il modello predittivo di PyRMD utilizzando i risultati del docking molecolare. Una volta

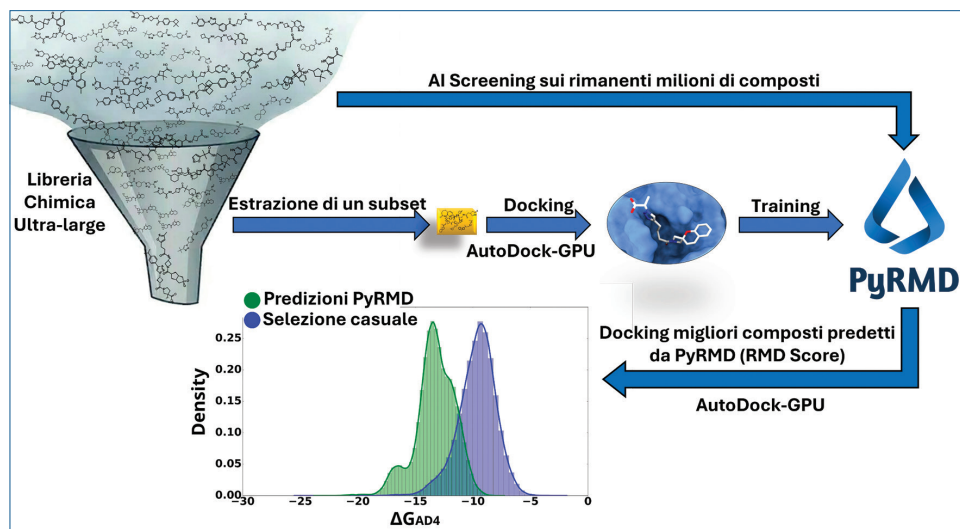


Fig. 3 - Diagramma di flusso del protocollo gerarchico PyRMD2Dock

addestrato, l'algoritmo opera una sistematica prioritizzazione della libreria, identificando i composti con la più alta probabilità statistica di affinità sulla base delle loro caratteristiche chimiche. I migliori candidati vengono infine sottoposti alla validazione strutturale mediante docking accelerato tramite AutoDock-GPU. Questa integrazione sequenziale permette di ottimizzare l'impiego delle risorse computazionali, limitando le simulazioni termodinamiche onerose ai soli composti prioritizzati dall'intelligenza artificiale.

Unificazione degli ambienti di calcolo: PyRMD Studio

L'adozione di strumenti computazionali avanzati è spesso ostacolata dalla ripida curva di apprendimento richiesta dalle interfacce a riga di comando e dalla necessità di configurare complessi ambienti di programmazione. PyRMD Studio nasce per rispondere a queste sfide in termini di accessibilità e standardizzazione. Si tratta di una suite integrata all'interno di un'interfaccia grafica (GUI) user-friendly che condensa i moduli LBVS (PyRMD), lo SBVS (PyRMD2Dock) e l'analisi dei risultati in un unico ambiente interattivo.

Dal punto di vista informatico, PyRMD Studio si affida al gestore di pacchetti Anaconda per facilitare la gestione dell'ambiente, garantendo che le versioni delle dipendenze e delle librerie scientifiche (come *pandas*, *scikit-learn*, *NumPy* e *matplotlib*) rimangano coerenti su diversi sistemi operativi.

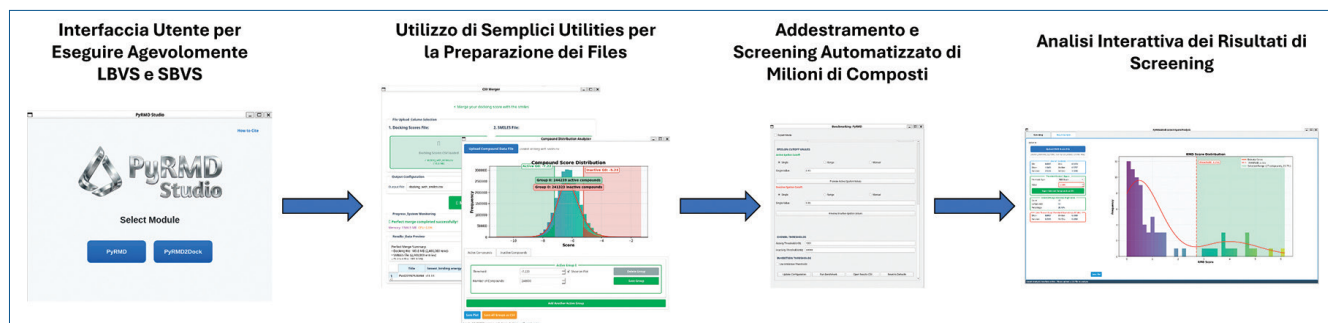


Fig. 4 - Panoramica dell'ecosistema PyRMD Studio

Per ottimizzare la fase iniziale di addestramento e validazione, la suite integra utility specifiche progettate appositamente per agevolare il compito dell'utente. Attraverso un'interfaccia semplice e interattiva, i ricercatori possono definire in modo dinamico diversi parametri (come le soglie di attività o i punteggi di docking dei vari dataset) e affinare il modello, personalizzando gli esperimenti di benchmarking al fine di ottenere le migliori prestazioni predittive possibili. Il rigore scientifico di questa fase è ulteriormente garantito dall'implementazione del clustering di Butina, impiegato per separare rigorosamente i cluster strutturali tra il training set e il test set, prevenendo il fenomeno del *data leakage* e mitigando l'*overfitting*. Infine, per la selezione dei composti a valle dello screening, la suite include uno strumento di analisi dei risultati per permettere all'utente di visualizzare dinamicamente la distribuzione dei risultati da lui ottenuti e applicare soglie visive, facilitando la separazione tra il rumore di fondo e i potenziali composti da avviare alla validazione sperimentale (Fig. 4). L'interfaccia grafica unificata democratizza l'accesso al VS avanzato, integrando in un unico ambiente i moduli LBVS e SBVS. La suite mette a disposizione dell'utente utility interattive per personalizzare dinamicamente la fase di benchmarking e ottimizzare i modelli predittivi utili per lo screening di milioni di molecole in tempi rapidi. Infine l'utilizzo di strumenti dedicati all'interno della suite accompagnano il ricercatore fino all'analisi visiva dei risultati per un'efficace selezione dei potenziali hit.

Validazione sperimentale: identificazione di un modulatore per GPR183

L'applicabilità e la solidità dell'infrastruttura computazionale descritta sono state provate sperimenta-

talmente nel contesto della modulazione dei recettori accoppiati a proteine G (GPCR). Lo studio si è focalizzato sul recettore orfano GPR183 (noto anche come EBI2), un recettore di membrana attivato da specifici oxisteroli endogeni (come il $7\alpha,25$ -diidrossicolesterolo) e profondamente implicato nella migrazione linfocitaria e nell'eziopatogenesi di disordini autoimmuni e infiammatori.

I GPCR spesso esibiscono un'attività basale intrinseca in assenza di agonisti. L'obiettivo terapeutico consisteva nell'identificazione di un agonista inverso, ovvero una molecola capace di legarsi al recettore stabilizzandone la conformazione inattiva e sopprimendone l'attività costitutiva.

Impiegando i protocolli potenziati dall'AI, è stata scansionata un'ampia libreria commerciale.

L'applicazione del protocollo PyRMD2Dock ha restituito un pannello ristretto di promettenti composti. La successiva validazione *in vitro* è stata condotta avvalendosi di biosensori conformazionali basati sul fenomeno della Bioluminescence Resonance Energy Transfer (BRET). Tali sensori consentono di monitorare in tempo reale e in cellule vive i riarrangiamenti strutturali del recettore indotti dal legame con il ligando.

La campagna ha condotto all'identificazione di un agonista inverso di nuova concezione strutturale, caratterizzato da un'elevata affinità micromolare. Più significativamente, l'integrazione dei dati sperimentali con l'analisi computazionale delle pose di legame ha permesso di delucidare un intricato meccanismo di "switch" molecolare all'interno del fascio transmembrana. Sono stati identificati specifici residui amminoacidici critici che fungono da interruttori per la transizione di fase del recettore GPR183, offrendo prospettive inedite per il design razionale di futuri agenti antinfiammatori (Fig. 5).

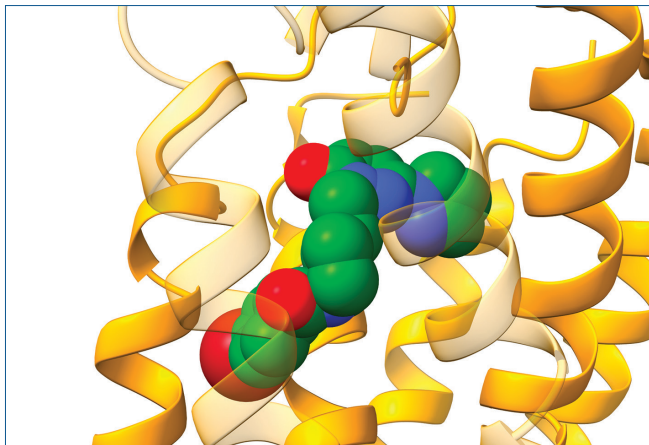


Fig. 5 - *Binding mode* del ligando maggiormente attivo (verde) scoperto mediante l'applicazione di PyRMD2Dock in complesso con la proteina GPR183 (arancione)

Conclusioni

L'integrazione dell'intelligenza artificiale nei flussi di lavoro della chimica computazionale sta determinando un cambio di paradigma nell'esplorazione dello spazio chimico. L'analisi congiunta dei lavori qui presentati evidenzia l'evoluzione e l'impatto pratico della suite basata sull'algoritmo Random Matrix Discriminant (RMD).

In primo luogo, l'implementazione automatizzata in PyRMD offre una soluzione robusta alle sfide del Ligand-Based Virtual Screening (LBVS). Il software ha dimostrato di azzerare la soggettività operativa legata alla preparazione dei dati e alla scelta dei composti inattivi, raggiungendo prestazioni predittive superiori a complesse architetture di deep learning (es. reti neurali basate su grafi), a fronte di un onere computazionale drasticamente inferiore. In secondo luogo, per affrontare la sfida computazionale posta dalle librerie chimiche "ultra-large" nell'ambito dello screening virtuale basato su struttura, il protocollo gerarchico PyRMD2Dock ha confermato l'efficienza di un approccio ibrido. L'impiego del modello AI combinato con il docking molecolare accelerato via hardware (AutoDock-GPU), permette di scansionare archivi con miliardi di molecole *make-on-demand*, arricchendo significativamente il tasso di identificazione (*hit rate*) e ottimizzando l'impiego delle risorse di calcolo.

Dal punto di vista dell'usabilità e della standardizzazione, l'ambiente containerizzato PyRMD Studio risolve il problema della riproducibilità tecnica. Fornendo un'interfaccia unificata, la suite svincola

il ricercatore dalle complessità della programmazione, integrando in un solo flusso di lavoro l'addestramento dei modelli, la preparazione dei dati e l'analisi avanzata post-screening.

Infine, la campagna di screening sul recettore GPR183 attesta l'applicabilità traslazionale di queste metodiche. La convergenza tra i protocolli predittivi e l'impiego di biosensori conformazionali (BRET) non solo ha portato all'isolamento di un agonista inverso caratterizzato da un chemotipo innovativo, ma ha consentito di mappare il meccanismo molecolare di attivazione recettoriale a livello dei domini transmembrana.

Complessivamente, l'adozione di protocolli computazionali integrati si conferma un elemento strutturale per il processo di *drug discovery*, accelerando l'identificazione di nuove entità chimiche e fornendo basi razionali per la comprensione di target biologici complessi.

BIBLIOGRAFIA DI RIFERIMENTO

- [1] J.L. Reymond, *Accounts of Chemical Research*, 2015, **48**, 722.
- [2] G. Amendola, S. Cosconati, *J. Chem. Inf. Model.*, 2021, **61**, 3135.
- [3] M. Roggia, B. Natale *et al.*, *J. Chem. Inf. Model.*, 2023, **63**, 5326.
- [4] L. Andersson, M. Roggia *et al.*, AI screening and a conformational biosensor reveal a GPR183 inverse agonist and a switch mechanism for receptor activation, (Draft Manoscritto), 2024.
- [5] B. Natale, M. Waqas *et al.*, PyRMD Studio: A Unified Suite for Next-Generation, AI-Powered Virtual Screening, (Draft Manoscritto), 2026.
- [6] S. Cosconati, Avogadro Colloquia, Roma, 2025.

AI and Chemical Space: the PyRMD Method

Artificial intelligence revolutionizes drug discovery by making an infinite chemical space accessible. This article describes PyRMD, an automated virtual screening tool based on the Random Matrix Discriminant algorithm. It illustrates the integration with docking (PyRMD2Dock), the PyRMD Studio suite, and the discovery of a modulator for the GPR183 receptor.